Lang Resources & Evaluation (2008) 42:41–73
DOI 10.1007/s10579-007-9049-1

Validation of spoken language resources: an overview of basic aspects

Henk van den Heuvel · Dorota Iskra · Eric Sanders · Folkert de Vriend

Published online: 12 December 2007

Abstract Spoken language resources (SLRs) are essential for both research and application development. In this article we clarify the concept of SLR validation. We define validation and how it differs from evaluation. Further, relevant principles of SLR validation are outlined. We argue that the best way to validate SLRs is to implement validation throughout SLR production and have it carried out by an external and experienced institute. We address which tasks should be carried out by the validation institute, and which not. Further, we list the basic issues that validation criteria for SLR should address. A standard validation protocol is shown, illustrating how validation can prove its value throughout the production phase in terms of pre-validation, full validation and pre-release validation.

Keywords Quality · Assessment · Validation · Evaluation · Language resources · Spoken language resources

Abbreviations

ASR	Automatic speech recognition
DTD	Document type definition
ELRA	European language resources association
IMDI	ISLE meta data initiative
LR	Language resource
OLAC	Open language archives community
POS	Part of speech
SLR	Spoken language resource

H. van den Heuvel (✉) · E. Sanders · F. de Vriend
SPEX/CLST, Radboud University Nijmegen, Nijmegen, The Netherlands
e-mail: H.vandenHeuvel@let.ru.nl

D. Iskra
LogicaCMG, Nieuwegein, The Netherlands

SPEX	Speech processing expertise centre
TTS	Text to speech
QQC	Quick quality check
QQC_DB	QQC on database
QQC_DF	QQC on description form
WLR	Written language resource

1 Introduction

Language resources (LRs) are essential for efficient and effective research and application development. To this end they should be of high quality. This makes *quality assessment* a key issue in LR production. Both terms *Quality* and *Assessment* need some definition in this context. Cieri (2006) argued that the quality of an LR cannot be expressed on a single dimension 'good–bad', but comprises multiple dimensions. We identify the principle dimensions of LR quality as:

- Consistency (both internal and with documentation)
- Suitability/usability for the need of the users
- Reusability/extensibility of the data
- Compliance with best practices
- Completeness and clarity of the documentation
- Validation by independent validator
- Accessibility

In this context *assessment* is the process of collecting valid and reliable information about an LR, integrating it, and interpreting it to make a judgement or a decision about its quality. Two approaches have been developed to assess the quality of LRs over the last decade: *evaluation* and *validation*.

Evaluation of an LR implies testing it by employing the LR in an actual application (Dybkjaer et al. 2007). An evaluation does not only require data sets but also tools/engines and scoring procedures/scripts for the application at hand. For that reason the result of the evaluation is dependent on (the quality of) both: data and engines, and one can evaluate either one or both. Evaluation commonly focuses on the quality of systems or system components, as in the NIST Spoken Language Technology Evaluations.[1] For such evaluations, new LRs are used that are not accessible before the evaluation; obviously the evaluation database is the same for every comparative test in the evaluation. Alternatively, LRs as such can be evaluated; to that end the performance of various LRs on the same system can be measured. For instance, the usefulness of a speech database with car recordings can be evaluated by showing that a speech recognition engine trained on this database performs better than the same system trained on another database that was not recorded in the car environment.

[1] http://www.nist.gov/speech/tests/index.htm

Validation refers to the other approach to assess the quality of the LR. Validation of an LR is defined as a check of an LR against its specifications, augmented by a set of tolerance margins for deviations of these specifications (Van den Heuvel et al. 2004b). For example, the specifications state that 50% of the speakers should be male, and for validation a deviation of 5% is permitted. The full set of specifications and tolerance margins are the validation criteria for an LR. Output of a validation is typically a report that lists all checks performed together with an account of the results of the checks. Validation does not involve application testing to judge the quality of the data.

Evaluation and validation are both essential means of quality assessment. Training, (development), and test databases should be properly validated before the evaluation can be sensibly conducted. Thus, "validation" and "evaluation" are quality assessment measures that are independent of and complimentary to each other.

This paper deals with the *validation* of LRs, more specifically of *spoken* language resources (SLRs). SLRs are annotated collections of speech data. The difference between a mere collection of speech and an actual SLR is "the fact that the latter is augmented with linguistic annotation (i.e. a symbolic representation of the speech)", as is attested in the EAGLES handbook (Gibbon et al. 1997, p. 146). On the other hand, collections of annotations without accompanying speech data cannot strictly be called SLRs, even when these annotations clearly refer to spoken versions of the database entries, as is the case for phonemic transcriptions.

The relevance of validation of large SLRs emerged when the SpeechDat project (Höge et al. 1997) was started around 1995. The SLRs within this project were produced in a European framework according to design and recording specifications similar to the American-English Macrophone corpus (Bernstein et al. 1994) and the Dutch Polyphone corpus (Den Os et al. 1995). The SpeechDat SLRs were, however, produced by a large consortium, the idea being that each consortium member would produce from one to three SLRs and obtain the SLRs produced by the other partners at the end of the project. The Speech Processing Expertise Centre (SPEX) was included in the consortium as the validation centre with the task of monitoring the quality of data and ensuring that all databases would be of comparable quality. Another objective of SpeechDat was that the SLRs would become available to third parties after the end of the project. This was another reason for involving an independent validation centre.

Since SpeechDat, SPEX has been involved as a validation centre in many projects, particularly in data collections supported by the EU, such as SpeechDat Car (Moreno et al. 2000a), SpeeCon (Iskra et al. 2002), and OrienTel (Iskra et al. 2004). The experience on SLR validation gained over the years has been reported at conferences, tutorials and summer schools. This paper presents a comprehensive and up-to-date overview of our experience in the field, more in particular of the relevant issues that according to us are important for the validation of SLRs (i.e. annotated speech corpora including lexicons for prompted speech recordings). Much of our expertise has been developed in close cooperation with ELRA and its validation committee.

In this paper we will address basics of validation (Sect. 2), relevant issues for defining validation criteria (Sect. 3), validation types and procedures (Sect. 4) and

will conclude with lessons learnt (Sect. 5). It should be noted that the paper does not intend to present or analyze a survey of errors that we came across as a validation centre. Our main purpose is to convey that validation is an essential element in the quality assessment and quality assurance of LRs, and to pinpoint the relevant issues involved in LR validation, more particularly in SLR validation.

2 Validation basics and principles

Basic aspects of SLR validation have been addressed previously in Van den Heuvel et al. (2000), Schiel and Draxler (2003), Van den Heuvel et al. (2004b). A brief overview of SLR validation is also presented by Maegaard et al. (2005). Most of the issues presented in this section are so general that they apply to other LRs as well.

2.1 Objectives

The result of an SLR validation is commonly a validation report. This report presents a systematic survey of the validation criteria and the degree to which they were met by the SLR. It can serve a variety of purposes:

1. Quality assurance: in this case the validation report attests that the SLR meets the minimum of required specifications;
2. Quality improvement: the validation report shows where the SLR can be improved by listing which of the validation criteria were not met.
3. Quality assessment: the validation report can be added as an appendix to the SLR itself, especially if remaining errors have not been corrected.

2.2 Strategies

SLR validation can be performed in two fundamentally different ways: (a) Quality assessment issues and checks are addressed in the specification phase of the SLR. That is, during the definition of the specifications the validation criteria are formulated, and during the recording process pre-validations on the data are carried out. (b) The production of an SLR is completed, and the validation criteria and procedure are defined (and carried out) afterwards. Furthermore, validation can be done either in house by the producer (internal validation) or by another organization (external validation). This is schematically shown in Table 1.

Internal pre-production validation (1) in this table is in fact essential for proper database production. Each LR producer is responsible for the database quality during collection and processing of the data. Internal post-production validation (2) should be an obvious part of this procedure. These principles are employed by the Linguistic Data Consortium (LDC) (Cieri and Liberman 2000; Strassel et al. 2003). The LDC has an independent validation team as part of their organization (Cieri, personal communication). External pre-production validation (3) is the preferred choice, if the production of a database is sub-contracted or if LR-production is

Table 1 Four types of validation strategies

Validator	Validation scheduling	
	During specification and production	After production
Internal	(1)	(2)
External	(3)	(4)

carried out in a consortium. Combined with external post-production validation (4), this strategy was adopted by many European Union (EU) funded projects, where all producers performed internal quality checks, whilst SPEX served as an independent external validation institute. SPEX was closely involved in the specifications and performing intermediate and final quality assessments. An overview of these projects is presented in Table 2. In this EU-context, all four validation strategies shown in Table 1 were carried out. This two-dimensional view of the SLR validation process is obviously valid for other types of LRs as well, cf. Fersøe (2004) for lexicons.

2.3 Approval authority

When the validation takes place internally, the approval authority is with the producer. This is not the case when the producer is not the owner of the SLR (e.g. production is sub-contracted), or when the SLR is produced within a consortium of partners producing similar SLRs with the aim of mutual exchange, as in SpeechDat. In these cases an external validation institute makes an objective assessment to ascertain whether a producing party has fulfilled the requirements set out by the patron/consortium. The tasks of the validation institute are then to check an SLR against the predefined validation criteria, and then to put a quality stamp on it after a successful check.

The owner (resp. consortium) should decide upon the acceptability of an SLR; the validation report serving as factual information basis for the decision. In SpeechDat like projects, however, the approval of an SLR is commonly done by voting. In these cases, the process is to send a validation report to the producer for comments. Minor textual or formatting errors that can be easily corrected have to be repaired in the SLR and clarifications for larger discrepancies included in the final report. The validation institute requests votes based on the finalized validation report. After voting the outcome is reported to all parties concerned.

2.4 Role of a validation institute

Validation is just one element in the process of quality control of SLRs. Repairing imperfections is the next stage. It is important to distinguish between the validation and correction of an SLR. The two tasks should not be performed by one and the same institute. A conflict of interest may arise when the validation institute is, in the end, checking its own corrections. The appropriate procedure is that the producer

Table 2 Overview of SLR collection projects with an external validation component

Project	Type of SLR	Number of SLRs	Period	Ref.
SpeechDat(M)	Fixed telephone network, for voice-driven teleservices, European languages	8	1994–1996	Höge et al. (1997)
SpeechDat(II)	Fixed and cellular telephone network, for voice-driven teleservices, European languages	28	1995–1998	Höge et al. (1999)
Speechdat-Car	Car recordings incl. GSM channel, European languages	9	1998–2001	Moreno et al. (2000a)
SpeechDat-East	Fixed telephone network, for voice-driven teleservices, Central and East European languages	5	1998–2000	Van den Heuvel et al. (2001)
SALA	Fixed telephone network, for for voice-driven teleservices, Latin America	5	1998–2000	Moreno et al. (2000b)
SALA II	Cellular telephone network, for for voice-driven teleservices, America (full continent)	16	2002–2005	Van den Heuvel et al. (2004a)
Speecon	Broadband recordings for commanding consumer devices (major world languages)	28	1999–2002	Iskra et al. (2002)
Network-DC	Broadcast News (Arabic)	1	2000–2001	http://www.elda.org/article45.html
OrienTel	Fixed and Mobile telephone network, for voice-driven teleservices (Oriental region)	23	2001–2003	Iskra et al. (2004)
TC-STAR	Parliamentary speeches and Text To Speech (TTS)	3	2004–2007	Van den Heuvel et al. (2006)
LILA	Mobile telephone network, for voice-driven teleservices (Asian and Pacific region)	5	2005	Moreno et al. (2004)

Information about all projects can be obtained via http://www.speechdat.org. For TC-STAR see: http://www.tc-star.org

corrects the deviations found and that the validation institute again checks the correctness of the adjustments.

The best situation is when the validation institute is involved from the very beginning of the design of SLRs. Throughout the design phase, the institute can contribute expertise towards defining and fine-tuning the specifications. It can also make clear from the start which of these specifications can be reliably checked by the institute. During the specification phase the validation institute is responsible for addressing the definition of the tolerance margins for deviations of the validation specifications. For example, if half of the recordings in an SLR of 2,000 speakers should come from male speakers, will the SLR still be acceptable if it contains 999 male speakers, or 975, or even fewer?

When the specifications have been agreed upon, the contribution of the validation institute can be of great value by carrying out quality checks at strategic points during the production process. In Sect. 4.1, a comprehensive scheme of quality controls throughout the production process is presented.

An important issue remains though: who checks the validator? When a preliminary version of the validation report is written, the provider has the first right to comment on the findings of the validation institute. It is in most cases possible to achieve a consensus. In cases where consensus cannot be achieved, the validation institute may decide to consult one or more other experts to check the disputed part of the data, and go back to the producer with the new results.

It is very important that the validation institute provides efficient feedback on data submissions, and keeps all communication channels open for consultation and feedback on the results found. In practice, this means that:

- The arrival of a data set at the validation office is reported to the producer instantaneously.
- The data set is immediately checked for readability and completeness in terms of required files. This is of major importance if the SLR cannot be validated straight away. Readability and completeness issues can be resolved by the provider while the SLR is awaiting its turn.
- If possible, the producer should be allowed to resubmit defective files on the fly during validation.
- The validation report is first reviewed by the producer before it is disclosed to anyone else. This is necessary to avoid and remove any misunderstandings in the text of the report. For instance, a reported error may in fact be a lack of clarity in the documentation, and should be repaired there, not in the database itself. Based on the producer's comments a final report is edited which can be distributed to other partners in the consortium. It can also be included as part of the SLR.

3 SLR validation: what and how

This section contains more practical information about the contents of an SLR that can be validated and how this can be done. There are a number of relevant

elements to be validated in an SLR, which are successively addressed in the next subsections.

1. Documentation
2. Database format
3. Design and contents
4. Acoustical quality of the speech files
5. Annotation files
6. Pronunciation lexicon
7. Speaker and environment distributions
8. Orthographic transcriptions

For each of these items we will list a number of basic considerations and typical validation criteria. These criteria were developed during discussions in many SLR production projects (see Table 2) in which both SLR producers and validation centre aimed to strike a balance between delivering high quality SLRs and safeguarding the feasibility of data collection in practice. A more detailed overview of validation criteria can be found in Schiel and Draxler (2003) and Van den Heuvel et al. (2000). For further illustration the appendix contains a full listing of validation criteria as used in the SALA-II-project. One can use this list as an example list for the validation of an SLR.

3.1 Documentation

An SLR is rarely self-explanatory. Therefore, a good documentation should accompany the SLR. The documentation should contain:

- An account of the specifications of the SLR;
- An account of how they were fulfilled;
- Instructions on how to use the SLR.

For a user the documentation is of paramount importance to obtain a view of the usability of the SLR for the intended application. Common practice is that the producer writes the documentation at the end of the production process and in a great hurry. Moreover, the producer knows exactly what is in the SLR. These circumstances may lead to a cryptic and incomplete documentation that is not helpful to a user. For that reason, the validation institute can provide a documentation template. This has a number of advantages:

- All relevant aspects to be documented are listed beforehand;
- The documentation is a proper reflection of the specifications of the SLR;
- All documentation files within a multi-SLR project have the same uniform structure;
- The subsequent validation of the documentation by the validation institute is facilitated.

The validation institute checks if all relevant aspects of an SLR (see the list in Sect. 3 above) are properly described in terms of the three C's: clarity, completeness and correctness.

The documentation is the fundamental source of information for a user. The SLR may contain treasures of potentials for specific applications, but if they are not properly documented the gems of these treasures will remain hidden. Both for a user and a validation institute the worst situation arises if the SLR itself has to be used in order to reverse-engineer the producer's intentions. Therefore, the documentation is more than just a component of the SLR, it is the very key to it.

Relevant validation criteria for the documentation:

The documentation should contain a clear, correct and complete description of:

- Owner and contact point.
- Database layout and media.
- Application potential for the SLR.
- Directory structure and file names.
- Recording equipment.
- Design and contents of the recordings.
- Coding and format of the speech files.
- Contents and format of the annotation files and speech files.
- Speaker demographic information.
- Recording environments distinguished.
- Transcription conventions and procedure.
- Lexicon: format and transcription conventions included.

3.2 Database format

The database format serves the accessibility of an SLR. For that reason it is important that files are present at the documented locations, and in the correct format. This is especially relevant in order to enable automatic search.

Relevant validation criteria for the format:

- Directory structure is as documented.
- File names are as documented.
- Empty (i.e. zero-length) files are not included.
- Each speech file is annotated (either in a corresponding annotation file or in a speech file header).
- Each annotation file is connected to an existing speech file and vice versa.
- The format is a well-known standard or it is well documented.
- The database is free of viruses.

3.3 Corpus design and contents

Design and content checks include quality measures at several levels. Validation of the SLR design comprises the test whether all types of speech material that are specified in the documentation are present in the SLR and in sufficient quantities.

For SLRs with prompted material, it is necessary to make sure that all data *types* (e.g. digits, application words, date and time words, names) as specified in the documentation are included in the prompts. At prompt level the SLR should be designed such that all types appear and with sufficient tokens (e.g. digits, application words, phonemes). The frequency of the tokens at prompt level can be regarded as the theoretical upper bound of the recordings. At the end of the production, fewer tokens will commonly be contained in the SLR. This may be due to skipped prompts, missed words in a recorded item, mispronounced or truncated words, or extreme line or background distortions. This is reflected at the transcription level. Therefore, the minimum number of tokens for an item (word, phoneme, digit) at the *prompt* level can be accompanied by another criterion for the minimum number of tokens required at the *transcription* level. This number is partly dependent on whether or not the recordings are supervised. In unsupervised recordings such as telephone calls, the practical experience is that 80–85% of the upper bound can be reasonably achieved. In supervised recordings a speaker can be stopped to repeat a mispronounced prompt and the threshold can be set to a higher percentage (90–95%).

For SLRs with unprompted material other content specifications, and thus other validation criteria, apply. For Broadcast news databases these will be directed towards type of broadcasts and topics, minimum hours of transcribed speech, permitted time period between the recordings. For human–human dialogues the design specifications will address type of dialogue (problem solving, information seeking, chat, etc.), the relation between the speakers, the topic(s) under discussion, the degree of formality, and the use of scenarios (if any). For human–machine dialogues, important design parameters are the domain(s) and topic(s) under discussion, the dialogue strategy followed by the machine (system-driven, mixed-initiative), the type of system (test, operational service), and the instruction to the speakers (if any).

Relevant validation criteria for the design and contents of (prompted) SLRs:

- All mandatory corpus items according to the documentation are included.
- Number of missing files per corpus item is less than XX%.
- At the transcription level about YY% of the theoretically possible tokens are present.

3.4 Acoustic quality of the speech files

It is not easy to find a bundle of acoustic features that can be processed automatically in order to obtain an impression of signal quality that equals the

impression of human judgement. The practical estimate SPEX is currently using is a combination of the average clipping rate, Signal-to-Noise Ratio (SNR), and mean sample value. Files, or portions of files, or groups of files, with excessive values on these parameters are selected for auditory inspection of signal quality. On the basis of the final human judgment it is decided if the files are acceptable. Also speech file duration can be used to validate sound quality. In SLRs with prompted material, extreme long or short durations of files can indicate serious recording defects.

Relevant validation criteria for the acoustic quality:

- Empty speech files are not permitted.
- Acoustic quality of the speech files is measured, based on:
  - Clipping rate
  - SNR
  - Mean amplitude
  - File duration

Apart from the above measurements, the speech files can be checked for a minimum period of silence at the beginning and/or end of the file.

For SLRs with short utterances stored in separate files, a good procedure is to compute the acoustic measures over the complete file and average the outcomes over all the files of a speaker/session. In this way corrupted sessions can be spotted. However for broadcast news SLRs or SLRs with speeches, the acoustic measurements should be made on a per file basis, excluding the untranscribed portions where background noise (e.g. music, commercials, applause) is present.

Whether or not 'bad' recordings should be discarded from the database is a controversial issue. On the one hand, evidently corrupt signals should be deleted. On the other hand, as much speech signal as possible should be retained; it is 'always good for something', e.g. as test material. Obviously, if recordings are intended for speech synthesis purposes, criteria for discarding distorted files are much stricter than for SLRs intended for training Automatic Speech Recognition (ASR) engines.

3.5 Annotation files

In most SLRs speech files come with accompanying annotation files containing the orthographic transcription of the speech file and some other information such as speaker properties, recording environment, and characteristics of file formats.

The formal part of annotation of meta-data is greatly pushed by standardization initiatives. Initiatives such as the International Standards for Language Engineering (ISLE) Meta Data Initiative (IMDI, http://www.mpi.nl/IMDI; Wittenburg et al. 2006) and Open Language Archives Community (OLAC, Simons and Bird 2003) pave the way for further formal validations of annotation schemes.

Annotation files are also referred to as label files. In SpeechDat-context they contain a label followed by the actual content information or transcription. The label files should obey the correct format. Ideally, they can be automatically parsed without yielding erroneous information.

With XML-encoded annotation files there is the possibility of providing producers with a form of remote validation where they are able to reference a Document Type Definition (DTD) or Schema file that enforces some of the formal characteristics of the annotations. For this no data has to be transported to the validation centre at all. The only data transported is the relatively small DTD or Schema file that resides on the web server of the validation institute (De Vriend and Maltese 2004). For the actual validation report the check is finally also performed by the validation institute itself.

Relevant validation criteria for the annotation/label files:

- No illegal labels are used.
- All label files contain legal values.
- Labels do not contain empty values (unless intended so).
- XML files are well formed and valid against DTD (if included).

3.6 Pronunciation lexicon

A pronunciation lexicon (if part of an SLR, or as an independent LR) can be checked both at a formal and at a content level. At the formal level the encoding and format of the lexicon is examined. At the content level the information contained in phonetic transcriptions and other lexical information is examined in terms of correctness. For content checks like these, it is common to employ native speakers, although near-native speakers could also accomplish the task very well. The main reason for restricting to native speakers is that near-nativeness is too vague a notion when one needs to reassure producers about the quality of the validations.

The validation of the phonetic correctness of the lexicon entries is typically organized as follows:

- A selection of 1,000 entries are randomly extracted from the lexicon;
- In case of pronunciation variants, only one variant of the phonetic transcriptions of an entry is checked;
- The check is carried out by a phonetically trained person who is a native speaker of the language in question;
- In case of multiple possible correct transcriptions, the transcription given by the producer receives the benefit of the doubt;
- The given transcription is correct if it represents a possible pronunciation of the word (which is not necessarily the most common);
- Each transcription is rated on a 3-point scale: OK; error with respect to a single phone (minor); numerous errors (major).

Our experience has shown that the maximum allowed number of incorrect transcriptions can be placed between 3 and 5%. Usually the criteria are set a bit higher (and thus stricter) for TTS purposes than for ASR purposes.

For a maximum error percentage of 5%, the 95%-confidence interval for a sample of 1,000 transcriptions is 3.6–6.4%. This means that the lexicon is rejected when the number of errors exceeds 6.4%.

In many lexicons the phonetic transcriptions are accompanied by POS-tags. For lexicons developed in the LC-STAR project a similar procedure as shown above for phonetic transcriptions was used to check the POS tags (cf. Shammass and Van den Heuvel 2002).

Relevant validation criteria for the pronunciation lexicon:
Formal:

- All phone symbols in a lexicon agree with the specified set.
- All documented phone symbols are used.
- All used phone symbols are documented.
- All words found in the orthographic transcriptions are present in the lexicon.
- All words in the lexicon have at least one phonetic transcription.

Content:

- A maximum of XX% of the entries may contain one erroneous phone symbol in the transcription of an entry.
- A maximum of YY% of the entries may contain more than one erroneous phone symbol in the transcription of an entry.

3.7 Speaker and environment distributions

The specifications have to make sure that the recorded speakers represent a fair sample of the population of interest in terms of (typically) gender, age and dialectal background. Also the recording environments should reflect the targeted applications. That is, one would not expect to have a TTS database recorded in a car driving on a highway.

Relevant validation criteria for the speaker and environment distributions:

- Distributions of speaker properties are in agreement with specification.
- The recording environments are in agreement with the specifications.

3.8 Orthographic transcriptions

Similar to the lexicon, the orthographic transcriptions can be checked at a formal and at a content level. An SLR can only be accepted if at the formal level the orthographic encoding is correct and if all symbolic representations for non-speech events are documented and used. At the content level it is required that the orthographic transcriptions (including those of the non-speech events) are a correct representation of what is audible in the speech signal.

For the content check, the validation is split into two parts. There is a validation for the transcriptions of the spoken part of each utterance, and there is another validation for the transcriptions of the annotations of the non-speech events. A native speaker of the language performs the check on the orthography of speech. The transcription validation of the non-speech annotations is not necessarily done by a native speaker of the language, but by someone experienced in listening to non-speech events and capable of deciding which non-speech events should be transcribed or not. The transcriptions are checked by listening to the corresponding speech files and by correcting the transcriptions if necessary. As a general rule, the submitted transcriptions always have the benefit of the doubt; only overt errors are marked.

Typically, a sample of 2,000 utterances (about 2 h of speech) is selected. This gives statistically reliable confidence intervals for errors at the utterance level. This confidence level is dependent on the size of the sample (not of the population, i.e. the size of the SLR itself). For a maximum error percentage of 5% the 95%-confidence interval for a sample of 2,000 transcriptions is 4–6%. This means that the orthographic transcriptions are rejected when the number of utterances containing errors exceeds 6%.

Two types of errors are distinguished:

1. Errors in the transcription of speech.
2. Errors in the transcription of non-speech events.

The procedure described above works adequately for SLRs that are item-based, such as the databases from the SpeechDat family. In such databases an item is one utterance, e.g. a number, a date, a name etc. Transcription errors should be counted per utterance since a transcription error directly affects the usability of the whole item. The total number of transcription errors is less interesting than the number of items that contain one or more transcription errors.

Relevant validation criteria for the quality of the orthographic transcriptions:
Formal:

- A max of XX% of the speech files may miss an orthographic transcription (absent or empty transcription files).
- All transcriptions for non-speech events are described in the documentation.

Content:

- Maximum number of transcription errors.
  - For speech a maximum of YY% of the validated utterances (=files) may contain a transcription error.
  - For non-speech events a maximum of ZZ% of the validated utterances (=files) may contain a transcription error.

For other types of databases, this procedure is less suited. For instance, broadcast news databases are not divided in equivalent items, but in segments with speech of a similar nature. This means that both the validation procedure and error metric should be revised. A common measure for this is the WER (Word Error Rate), for

which a maximum of 0.5% can be demanded for speech and 1.5% for non-speech, for most types of SLRs. This is the case in the TC-STAR project (Van den Heuvel et al. 2006).

3.9 Automatic, manual or both?

Part of the validation can be done automatically. Apart from time saving, an automatic procedure provides a consistent level of precision that only a computer can offer. As a general rule the formal aspects of an SLR can be validated by scripts, and the content checks need human intervention.

Automatic checks are fast, consistent and can deal with large amounts of data (in fact with the full SLR), whereas manual checks are much slower but necessary where the checks focus on content, require expert knowledge, and are more aimed at empirical quality.

Since the production of a script is human labour and time-consuming, one should always consider if the automation of a check is time-effective. Writing and testing scripts and programs is mainly advantageous if large amounts of data have to be processed and/or if (many) more SLRs of the same type are expected for validation. Of course, the output of the scripts, in terms of reported errors should again be interpreted and reported by means of human labour and intervention.

On the other hand, evident manual work can be facilitated by scripts preparing the material and by interfaces that make manual verification work fast, efficient and less error-prone. For instance, for checking the orthographic transcriptions, a tool that quickly navigates through the selected material with simple buttons to indicate (types of) errors can seriously reduce the work load of the validator.

For the quality checks that were dealt with in the previous subsection, Table 3 shows a scheme of which checks are in general performed automatically and/or manually.

Table 3 Overview of manual and/or automatic validation work

Automatic	Manual (hand, ear)
	Documentation
Format/structure of SLR	
Design	
Speech files	Speech files
Annotation files	
Lexicon	Lexicon contents
Speaker and environment distributions	
Orthographic transcriptions (format)	Orthographic transcriptions (content)
	Interpretation of the output of the validation software

4 Validation types and procedures

The checks mentioned in the previous section are all carried out by the validation institute upon completion of SLR recordings, annotations and packaging. However, effective and efficient quality monitoring can be added at other points in the production process to ensure optimal quality. SPEX has developed a standard validation protocol in the course of its experience as a validation institute. This will be addressed in Sect. 4.1. Apart from that, other data quality monitoring services have been developed in other contexts; these will be addressed in Sect. 4.2.

4.1 Standard protocol

Over the years SPEX has developed a standard validation protocol for SLRs in SpeechDat-like projects, which is, apart from details, also applicable to other types of LRs. The protocol follows the steps outlined in Fig. 1. Clearly, the procedure reflects the current state of affairs and is open to further development and refinement.

4.1.1 Pre-validation

Pre-validation of an SLR is carried out before the stage of extensive data collection is entered. The main objective of pre-validation is to detect design errors before serious data collection starts. Secondary objectives are:

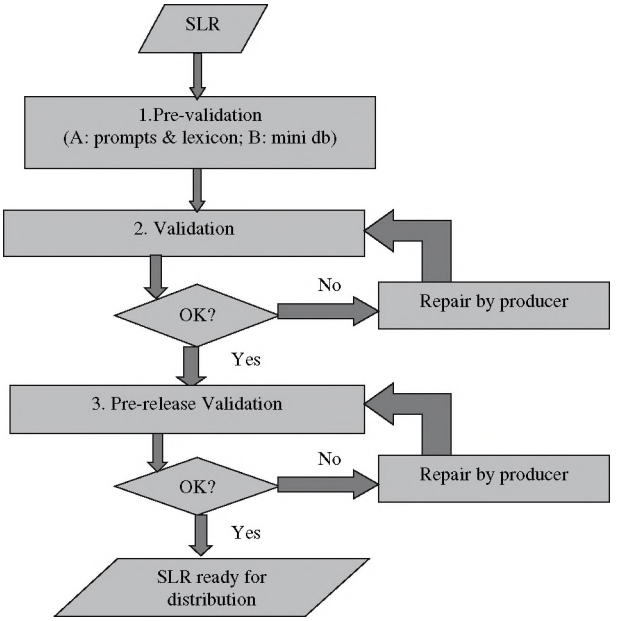

Fig. 1 Flow diagram depicting a standard validation protocol

- To enable the producer to go through the whole stage of documenting and packaging very early so that ambiguity and errors at the end are avoided.
- To enable the validation institute to develop and fine-tune software for validation of the full database.

At the pre-validation phase three components are assessed: prompt sheets, lexicon, and a mini database. The producer can deliver these components together as one package, or one-by-one, submitting a new component after the previous has been validated. Since pre-validation is diagnostic in nature, normally there are no iterations of repairs and new pre-validations.

Prompt sheet validation Before embarking on recording the speakers, the producers design prompt sheets. These prompt sheets should be an ideal representation of the content of the corpus items and the number of repetitions for each item. Since in practice not all intended material is recorded due to problems with the recording platform, or speakers omitting certain items altogether, not reading them correctly, stuttering or speaking in an environment with high background noise, etc., the reading scripts contain the (theoretical) upper bounds of types and tokens of what is achievable in a database.

The validation of the prompt sheets comprises checks with regard to the presence of the corpus items, adherence of their design to the specifications as well as the number of repetitions at word or sentence level calculated for the complete database. For phonetically rich words and sentences, if included, it can also be checked if a fixed minimum number of tokens per phoneme can be collected, provided that a lexicon containing all the words and their phonetic transcriptions is delivered as well.

If at this stage the prompt sheets do not fulfill the validation criteria (the absolute minimum which is required in the end), measures can still be easily taken to repair the errors. SLR producers indicate that they highly appreciate this part of validation which allows them to spot and repair errors in an early design stage. The prompt sheet validation is also a test for the specifications as it reveals parts which are underspecified and need further clarification.

Lexicon validation A formal check of the lexicon with regard to the format and the use of legal phoneme symbols is part of all the validation stages and can be carried out by the validation institute itself. However, the quality of the phonetic transcriptions has to be checked as well. Since this work needs to be done by phoneticians familiar with each language, the validation institute contracts this task to external experts. These experts obtain the relevant parts of the documentation describing the principles of the phonetic transcriptions employed by the producer. The experts obtain a sample (normally 1,000 entries) of the entire lexicon which they have to check manually. They are instructed to give the provided pronunciation the benefit of the doubt and only to mark transcriptions that reflect an overtly wrong pronunciation. This is in order to prevent marking as errors differences which are due to different phonetic theories or different ideas about what the 'most common' or 'best' pronunciation is.

Mini database validation Commonly, about 10 initial recordings are made in different environments and annotated. The data is formatted and packaged as if it

were a completed SLR, including documentation, and submitted to the validation institute. The purpose of this part of the pre-validation is to check if all items as specified in the prompt sheets are recorded and, if relevant, in the correct order. Further, the format, and the annotations are inspected, all with the aim of preventing errors during large-scale production. Since the documentation is included as well, the producers are forced to start documenting at an early stage. The advantages of this are clearly gained in the final production phase; the burden of documenting in that phase is greatly reduced to some final text editing and modifications of numeric tables.

4.1.2 Full validation

When all recordings are collected and annotated, the database is packaged and shipped to the validation institute for full validation. The purpose of the full validation is a quality assessment of the end product. At full validation, all checks as mentioned in Sect. 3 are carried out.

The validation institute may have a queue of SLRs to be validated. Because SLRs are typically handled in the order received, the validation institute performs a Quick Check. This is a quick formal test running the validation scripts to find out if all required files are included in the SLR and if they have the correct formal structure. If not, the producer is requested to submit updated versions of defective or missing files before actual validation takes place. Quick Checks allow the producer and the validation institute to work efficiently in parallel.

Since the validation of the (orthographic) transcriptions is restricted to a sample of all recordings, not all speech data is needed during full validation. For large SLRs such as those collected in SpeeCon, copying of all speech files onto a hard disk would use up the main part of the validation effort. For this reason, in SpeeCon and similar projects, the validation institute selected a list of 2,000 items during the Quick Check, for which the producer instantly had to provide speech files. Note that all orthographic transcriptions are already delivered for the quick check and that updates of the transcriptions are not accepted at a later stage. This is to avoid new transcriptions being made for the subset of files selected for validation.

If substantial shortcomings are found during validation, rectification and a subsequent re-validation of an SLR may become necessary. This is decided by the owner or the consortium in charge of the SLR production. Since usually not all parts are defective, re-validation is normally of a partial nature. Re-validations may iterate until approval of the SLR is achieved.

4.1.3 Pre-release validation

The validation of a complete database results in a report containing a list of errors which were found in the database. Some of them are irreparable and related to flaws in the (manual) annotation and/or the design of the database or the recordings themselves. However, a large number are usually minor and refer to the

documentation, label files or other text files which are produced during post-processing. These errors can easily be repaired and the producers are willing to do that. The danger, however, is the introduction of new errors or format inconsistencies during repair. Therefore, a pre-release validation has been introduced so that the envisaged master disks can be checked again by the validation institute. The purpose of this validation is to make sure that the reparable errors which were found during complete validation have been fixed and that no new errors have been introduced.

After full validation the documentation file is augmented with an additional section: "Modifications after validation". It is checked if all changes agreed upon are included in this section and if they have been implemented in the submitted pre-release version. The validation software is run, so that all formal checks on the data are carried out once more.

If the pre-release validation is finished with a positive result, the database is ready for distribution and the producers are not allowed to make any more changes, however minor, since these corrections can introduce new (and larger) errors. The pre-release phase may have one or more iterations until the SLR is approved for distribution.

4.2 Other types and procedures

As the European Language Resources Association's validation unit for SLR, SPEX has worked together with ELRA's Validation Committee (Van den Heuvel et al. 2003) to establish additional means for SLR quality control.

The first instrument is the Quick Quality Check (QQC) (Van den Heuvel 2004). This is a brief validation concentrating on the formal aspects of an SLR. It is intended for SLRs that are already in ELRA's catalogue and for all SLRs that are about to enter it. The goal is first to obtain a gross idea of the (formal) quality of an SLR, and second, if the QQC indicates so, to mark SLRs for a more detailed validation.

The following principles have been adopted for the QQCs:

A. The QQC mainly checks the database contents against minimal requirements. These requirements are of a formal surface nature which enables a quick check. Content checks are included in other types of validations. Minimal requirements are formulated for a limited set of application domains: ASR, Phonetic Lexicons, Speech Synthesis. For each of the domains a template document for QQC is made.
B. Generally, a QQC should take about 6–7 h work at maximum.
C. For each SLR two QQC reports are produced: One for the provider and users on the quality of the SLR proper (QQC_DB); one for ELRA on the quality of the information on the description forms (QQC_DF). A description form is a brief data sheet containing the main properties of an SLR.

During the QQC_DB the SLR is checked for compliance with a set of minimal requirements and for correspondence with its own documentation. The QQC_DB

report is intended for ELRA's database users if the SLR is already in the catalogue and for the database providers if the database is new and not in the catalogue yet. Each QQC_DB report is sent to the SLR provider for comments. Based on these a new version of the QQC_DB report and/or of the SLR may result. With permission of the provider the QQC_DB report is made available through ELRA's catalogue on the web.

Each database at ELRA is accompanied by one or two description forms: a general description form and/or a specific description form (depending on the type of resource). These description forms contain the basic information about a database according to ELRA. The description forms are filled out in cooperation with the SLR provider. The form is used to inform potential customers about the database. The information provided in the description form should be correct. The correctness of this information is also a minimum requirement for a database and checked at the QQC. The QQC_DF report contains a quality assessment of the correctness of the information in the description forms.

A second means of monitoring and improving SLR quality is a bug report service (Van den Heuvel et al. 2002). This service is implemented and maintained at ELRA's website (http://www.elra.info). The idea is that errors in SLRs distributed by ELRA are reported by SLR users through this bug report service. An error list per SLR is maintained and attached to the SLR information in ELRA's catalogue on their website. This document contains a formal list of verified errors (Formal Error List, FEL). Patches or new SLR versions can be made to correct errors.

The access to the FEL through the web is free and allows bug reporting users to see the status. Based on an update of the FEL the provider of that SLR is asked to correct the erroneous SLR part. ELRA sends the corrected part to SPEX. If the provider cannot repair the incorrect files, ELRA or other institutions selected by ELRA produce the corrected part. SPEX checks that corrections are properly made and that the patch is as intended. These services have successfully been implemented for SLR, and similar services are now under development for Written Language Resources (Fersøe and Monachini 2004).

Finally, validation manuals have been written both for SLRs and WLRs. They are available from ELRA's URL. These documents describe validation guidelines, procedures and criteria that should be taken into consideration by providers of new LRs. The documents give an idea of how validation at ELRA takes place, and allow producers to anticipate relevant quality checks before delivery to ELRA. In this way, this also contributes to the improvement of SLR quality.

5 Concluding remarks

In this article we have clarified the concept of SLR validation. We have addressed the concepts 'quality' and 'assessment' and have elaborated on the different roles of 'validation' and 'evaluation' in quality assessment. Furthermore, we have presented basic principles in LR validation. We have pinpointed a number of relevant issues for defining validation criteria for SLR. A standard validation protocol has been shown illustrating how validation can prove its value all along the production phase

in terms of pre-validation, full validation and pre-release validation. Other relevant LR quality control instruments have been briefly presented, too.

From our experience as a validation centre in many (mainly European) projects we have learnt a number of valuable lessons:

- External validation is an important quality safeguard.
- If the validation institute is involved during the specification phase of an SLR, it can advise on the specification of the design and the formulation of the validation criteria.
- The validation institute can provide important input at strategic points along the data collection and annotation, not only after the completion of the SLR. A good pre-validation procedure can avoid mistakes that would not be reparable at the end.
- The validation institute needs to keep open communication channels to the SLR provider.
- Clear validation protocols help structuring the work and effective quality control.
- A documentation template provided by the validation institute is to the benefit of all involved parties (provider, validation institute and future users).
- A relevant part of the work of the validation institute is to find a proper balance between developing automatic checks by scripts and hand labour.
- The validation institute, as a rule, does not claim the approval authority for an SLR.
- The validation institute, as a rule, does not perform any of the required corrections itself to avoid the situation in which it is checking its own work.

Acknowledgements The authors would like to thank the anonymous reviewers who greatly helped with their valuable comments on text and contents of this paper.

Appendix: Validation criteria used in the SALA II project

1. Documentation

- File DESIGN.DOC is present
- Language of doc file: English
- Contact person: name, address, affiliation
- Description of number of CDs and contents per CD
- The directory structure of the CDs
  - database, block and session orderings
  - directories DOC, INDEX, TABLE (and optionally PROMPT, SOURCE)
- The format of the speech files (A-law, Mu-law, 8 bit, 8 kHz, uncompressed)

- File nomenclature
  - root files
  - names of speech files and label files
  - files in directories DOC, INDEX, TABLE (and optionally PROMPT, SOURCE)
- Contents and format of the label files
  - clarification of attributes (three letter mnemonics)
  - example of labelfile
- Description of recording platform
- Explanation of speaker recruitment
- Prompting information
  - connection of sheet items to item numbers on CD
  - sheet example
  - items must be spread over the sheet to prevent list effects (e.g. three yes/no questions immediately after another are not allowed)
- Description of all recorded items
- Analysis of frequency of occurrence of the phones represented in the phonetically rich sentences and phon. rich words at transcription level (format: table)
- Analysis of frequency of occurrence of the phones represented in the full database at transcription level (format: table)
- Transcription conventions
  - procedure
  - quality assurance
  - character set used for annotation (transcription) (ISO-8859)
  - annotations symbols for non-speech acoustic events must be mentioned Filled Pause, Speaker Noise, Stationary Noise, Intermittent Noise, Beep Tone
  - list of symbols used to denote word truncations, mispronunciations, distortion due to the cellular network transmission, and not understandable speech
  - case sensitivity of transcriptions
  - use of punctuation
- Lexicon information
  - procedures to obtain phonemic forms from orthographic input (lexicon generation and lay out)
  - splitting of entries only at spaces
  - (Reference to) SAMPA symbols used
  - case sensitivity of entries (matching the transcriptions)
- Speaker demographics
  - which regions, how many of each
  - motivation for selection of regions
  - which age groups, how many of each
  - sexes: males, females, also children?; how many of each
  - how many sessions by how many speakers

- Recording conditions:
  - description of recording environments
  - number of speakers per environment
- Information on test (set) specification
- The validation report made by SPEX (VALREP.DOC) is referred to

2. Database structure, contents and file names

- Directory/subdirectory conventions
  Format of directory tree should be
  \<database>\<block>\<session>
  - database: defined as <name><#><language code><name> is MOBIL
    <#> is 4 for SALA
    <language_code> is the ISO two-letter code for the language
  - block: defined as BLOCK <nn> where <nn> is a progressive number from 00 to 99. Block numbers are unique over all CD's. They correspond to the first two digits of <nnnn> below.
  - session: defined as SES <nnnn> where <nnnn> is the session code also appearing in file name
- File naming conventions
  All file names should obey the following pattern: DDNNNNCC.LLF
  - DD: database identification code
    For SALA II: B4 = cellular net
  - NNNN: session code 0000 to 9999
  - CC: item code; first character is item type identifier, second character is item number
  - LL: language code (as specified in Technical Annex)
  - F: speech file type
    A is for A-law; U is for Mu-law;
    O is for Orthographic label file
- NNNN in filenames is not in conflict with BLOCK and SES numbers in pathname
- Contents lowest level subdirectories should be of one call only
- All text files should be in MS-DOS format (<CR><LF>) at line ends
- A README.TXT file should be in the root describing all (documentation) files on the CD-ROM
- A file containing a shortened version of the volume name (11 chars max.) should be in the root directory. The name of this file is DISK.ID. This file supplies the volume label to UNIX systems that cannot read the physical volume label. Example of contents: MOBIL4EV_01
- A copyright statement should be present in the file COPYRIGH.TXT (root)
- Documentation should be in \<database_name>\DOC
  - DESIGN.DOC
  - TRANSCRP.DOC (optional)

  - SPELLALT.DOC (optional)
  - SAMPALEX.PS
  - ISO8859<nr>.PS
  - SUMMARY.TXT
  - SAMPSTAT.TXT
- Tables should be in \<database_name>\TABLE
  - SPEAKER.TBL (optional)
  - LEXICON.TBL
  - REC_COND.TBL (optional)
  - SESSION.TBL
- Index files (optional) should be in \<database_name>\INDEX.
  Mandatory are:
  - CONTENTS.LST
  - B4TST<language code>.SES
- Prompt sheet files (optional) should be in \<database_name>\PROMPT
- All sessions indicated in the documentation SUMMARY.TXT are present on the CDs
- Empty (i.e. zero-length) files are not permitted
- File match: For each label file there must be one speech file and vice versa
- Part of the corpus is designed for training and a smaller part for testing:
  - For databases of 1,000 sessions 200 test sessions are required, for databases with more than 2,000 sessions 500 test sessions should be defined.
  - No overlap between train and test sessions is allowed.
- All table files, and index files should report the field names as the first row in the files using tabs as in the data records following.
- The contents of the database as given in CONTENTS.LST should comprise:
  - CD-ROM volume name (VOL:)
  - Full pathname (DIR:)
  - Speech file name (SRC:)
  - Corpus code (CCD:)
  - Speaker code (SCD:)
  - Speaker sex (SEX:)
  - Speaker age (AGE:)
  - Speaker accent (ACC:)
  - Orthographic transcription of uttered item (LBO:)
  - The first line should be a header specifying the information in each record.
  - This file must be supplied as an ASCII TAB delimited file.
- The contents of the SUMMARY.TXT files should comprise:
  - The full directory name where speech and label files are to be found (DIR:)
  - the session number (SES:)
  - a string of typically N codes. Each item present is represented by its code. If the item is missing, a '- -' should appear.
  - recording date (RED:)
  - recording time of first item (RET:)
  - optional comment text

  - all these fields are separated by spaces
    Note: The contents of the SUMMARY.TXT file are not CD-dependent.
- Missing items per session
  Check with documentation (SUMMARY.TXT)
- The database should be free of viruses

3. Items

Check on mandatory corpus items

- 6 common application words (code A1-6)
  - read
  - set of 25–30 should be used, 25 of which are fixed for all
  - minimum number of examples of each word = #sessions/8 (at transcription level)
- 2 isolated digits (code I1-2)
  - read or prompted
- 1 sequence of 10 isolated digits (code B1)
  - each sequence must include all digits
  - optional are hash and star
- 4 connected digits (code C1-4)
  - 5+ digit number to identify the prompt sheet (optional) (C1)
    - read
- 9–11 digit telephone number (C2)
  - read
  - local numbers
  - inclusion of at least 50% cellular telephone numbers mandatory
- 16 digit credit card number (C3)
  - read
  - set of 150
  - if there is a checksum then formula must be provided
- 6 digit PIN code (C4)
  - read
  - set of 150
- ~30 digits per session are required
- digits must appear numerically on the sheet, not as words
- 1 date (code D1)
  - spontaneous
- 1 date (code D2)
  - read, wordstyle
  - analogue form
  - covering all weekdays and months, ordinals and year expressions (also exceeding 2000)
- 1 general or relative date (code D3)

  - read
  - analogue
  - should include forms such as TODAY, TOMORROW, THE DAY AFTER TOMORROW, THE NEXT DAY, THE~DAY AFTER THAT, NEXT WEEK, GOOD FRIDAY, EASTER MONDAY, etc.
- 1 application word phrase (code E1)
  - application word is embedded in phrase
  - read or spontaneous
  - at least five different phrases are required for each application word
  - a length of minimal three words per sentence is required
- 3 spelled words (code L1-3)
  - L1 is spontaneous name spelling linked to O1 (or to another item explicitly documented)
  - others are read
  - equal balance of all vocabulary letters
    artificial words can be used to enforce this balance
  - average length at least 7 letters
  - may include names, cities and other frequently spelled items
  - should primarily include equivalents of: A–Z, accent words, DOUBLE, APOSTROPHE, HYPHEN
- 1 money amounts (code M1)
  - read
  - currency words should be included
  - mixture of small amounts including decimals and large amounts not including decimals
- 1 natural number (code N1)
  - read
  - provided as numbers (numerically)
  - decimal numbers are only allowed for additional natural numbers
  - numbers should all be smaller than 1,000,000
- 6 directory assistance names (code O1-7)
  - 1 spontaneous name (e.g. forename) (O1)
  - 1 spontaneous city name (O2)
  - 1 read city name (list of at least 500 most frequent) (O3)
  - 1 read company/agency name (list of at least 500 most frequent) (O5)
  - 1 read proper name, fore- and surname (O7)
    (list of 150 names: both male and female names)
- 2 yes/no questions (code Q1-2)
  - spontaneous, not prompted
  - one question should elicit (predominantly) 'no' answers; the other (predominantly) 'yes' answers
  - also fuzzy answers should be envisaged
- 9 phonetically rich sentences (code S1-9)
  - read
  - minimum number of phone examples = #sessions/10
    - at transcription level

  - exception: rare phonemes:
    - these appear mainly in loan words AND
    - a max. of 10% of all phonemes in the language may be rare
  - each sentence may appear a max. of 10 times at prompt level
- 1 time of day (code T1)
  - spontaneous
- 1 time phrase (code T2)
  - read
  - analogue form
  - equal balance of all words
  - should include equivalents of: AM/ PM, HALF/QUARTER, PAST/TO, NOON, MIDNIGHT, MORNING, AFTERNOON, EVENING, NIGHT, TODAY, YESTERDAY, TOMORROW
- 4 phonetically rich words (code W1-4)
  - read
  - minimum number of phone examples = #sessions/10
  - at transcription level
  - exception: rare phonemes:
    - these appear mainly in loan words AND
    - a max. of 10% of all phonemes in the language may be rare
- each word may appear a max. of five times at prompt level
- Any additional, optional material:

Checks on presence of corpus files

The following completeness checks are performed:

Structurally missing corpus items:

- Which items are not recorded at all?

Incidentally missing files:

a. files that are not there
b. files with empty transcriptions in the LBO label field (effectively missing files)
c. corrupted speech files
d. files containing truncation and mispronunciation marks

SALA II has the following criteria for missing items:

- A maximum of 5% of the files of each mandatory item (corpus code) may be effectively missing.
- As missing files are counted: absent files, and files containing non-speech events only.
- For the phon. rich sentences a maximum of 10% of the files may be effectively missing or corrupted.

– There will be no further comparison of prompt and transcription text in order to decide if a file is effectively missing.
  As a consequence: If there is some speech in the transcription, then the file will NOT be considered missing, even if it is in fact useless.

4. Sampled data files

Coding

- A-law or Mu-law, 8 bit, 8 kHz, no compression

Sample distribution

File length:
We calculated the length of the files in seconds in order to trace spurious recordings if files were of extraordinary length.

Duration distribution over calls/ directories:

Length (s)	#Occurrences

Min–max samples:
We provide a histogram with clipping ratios. The clipping ratio is defined as the proportion of samples in a file that is equal to the maximum/ minimum value, divided by all samples in the file.

The histogram, then, is an overview of how many files were found in a set of clipping rate intervals.

Clip distribution over calls/directories:

Clipping rate (in %)	Occurrences

Mean values:
We computed the mean sample value of each item in each call. We provide a histogram with mean values below. The histogram, then, is an overview of how many files were found in a set of mean sample value intervals. This overview can be used to trace files with large DC-offsets.

Mean distribution over calls/directories:

Mean	Occurrences

Signal to Noise Ratio:
We split each signal file into contiguous windows of 10 ms and computed the Mean Square (energy) in each window. The mean sample value over the complete file was subtracted from each individual sample value before MS was computed. 30% of the windows that contained the lowest energy were assumed to contain line noise. In this way the signal to noise ratio could be calculated for each file by dividing the

mean energy over all windows by the mean energy of the 30% sample mentioned above. The result was multiplied by 10*log for scaling.

SNR distribution over calls/directories:

SNR occurrences

5. Annotation file

- Each line must be delimited by <CR><LF>
- No illegal SAM mnemonics used
- There are no SAM mnemonics missing
- All files must contain the same mnemonics. This holds as well for the optional mnemonics.
- No illegal field values should appear
- For spontaneous speech LBR should contain the specified identification word.

6. Lexicon

- Check lexicon existence (LEXICON.TBL)
- The entries should be alphabetically ordered
- Used SAMPA symbols are provided in SAMPALEX.PS
- In transcriptions only SAMPA symbols are allowed
- All SAMPA phoneme symbols should be covered
- Phoneme symbols must be separated by blanks
- A line in the lexicon should have the following format <grapheme form><TAB>[<frequency><TAB>]<phoneme transcription>[<altern.>][TAB] is ASCII 9.
- Each line is delimited by <CR><LF>
- All entries should have at least one phone transcription
- Alternative transcriptions are optional.
  They may follow the first transcription, separated by [TAB] or have a separate entry (only in case also frequency information is supplied)
- Orthographic entries are taken from the LBO-transcriptions from the label files. These LBO-transcriptions are as a rule split by spaces only, not by apostrophes, and not by hyphens.
- Words appearing only with * or ~ or % should not appear in the lexicon
- The lexicon should be complete
  - Check for undercompleteness (are all words in lexicon)
  - Check for overcompleteness
    (Undercompleteness is worse than overcompleteness. Overcompleteness cannot be a reason for rejection)
- Lexicon contents should be taken from actual utterances (from LBO), so the entries should exactly match the transcriptions.

- Optional information: stress, word/morphological/syllabic boundaries. But, if provided, then it should follow the SpeechDat conventions.

7. Speakers

- Obligatory information in the (optional) SPEAKER.TBL:
  - unique number (speaker/caller) SCD
  - sex SEX
  - age AGE
  - accent ACC
- Optional information:
  - height HET
  - weight WET
  - native language NLN
  - ethnic group ETH
  - education level EDL
  - smoking habits SMK
  - pathologies PTH
  - socio-economic status SOC
  - health HLT
  - tiredness TRD
- Each speaker only calls once. There is a tolerance of 5% of the speakers who may call twice.
- Balance of sexes
  - How many males, how many females, should match specification in documentation file
  - Misbalance may not exceed 5% (Each sex must be represented between 45 and 55% of the sessions)
- Balance of dialect regions
  - which dialect regions and how many of each should match specification in documentation file
  - ACC is used to check dialect balance, according to motivation in DESIGN.DOC
  - At least #sessions/20 speakers per dialect should be included
- Balance of ages
  - which age groups and how many of each should match specification in documentation file
  - Criteria
  - <16: >= 1% of speakers strongly recommended
  - 16–30: >= 20% of speakers mandatory
  - 31–45: >= 20% of speakers mandatory
  - 46–60: >= 15% of speakers mandatory
  - (The age criteria are meant for the whole database; they are not to be applied for male and female speakers separately)

8. Recording conditions

- Obligatory attributes of the (optional) REC_COND.TBL file should all be present and complete
- Obligatory attributes of the SESSION.TBL should all be present and complete
- The recordings are distributed as follows (check ENV):

Environment	Full database distribution	Each dialect region distribution
1. Car, train, bus	20 ± 5%	
2. Public place	25 ± 5%	≥20%
3. Street	25 ± 5%	
4. Home/Office	25 ± 5%	≥20%
5. Car kit (hand free mode)	5 ± 1%	No restriction

  - In each dialect at least 20% of the speakers are recorded in environments 1–3
  - In each dialect at least 20% of the speakers are recorded in the home/office environment
- Recordings from the fixed net are not included

9. Transcription

Validation by software tools

- Transliterations is case-sensitive unless specified otherwise.
  (In general lower case is used also at sentence beginning Only exception: proper names and spelled words, ZIP codes, acronyms and abbreviations. In the latter case blanks should be used in between the letters.)
- Punctuation marks should not be used in the transliterations
- Digits must appear in full orthographic form
- In principle only the following symbols are allowed to indicate non-speech acoustic events: [fil] [spk] [sta] [int] [dit]
  Other symbols (and language equivalents) must be mentioned in the documentation
- Asterisks should be used to indicate mispronunciations
- Double asterisks should be used for not understandable parts
- Tildes should be used to indicate truncations
- Percent signs should be used to indicate speech distortions due to transmission characteristics of the cellular network

Validation by a native speaker of the language

This validation was carried out by taking 1,000 short items and 1,000 long items.

The transcriptions in the label files for these samples were checked by listening to the corresponding speech files and correcting the transcription if necessary. In case of doubt nothing was corrected.

This check was performed by a native speaker of the language. The background noise markers were checked by a trained (non-native) validator.

- The evaluation comprised the following guidelines:
  - Two types of errors were distinguished: speech and non-speech transcription errors
  - Non-speech refers to [fil] [spk] [sta] [int] only
  - For non-speech all symbols were mapped to one during validation. i.e. If a non-speech symbol was at the proper location then it was validated as correct (regardless if it was the correct non-speech symbol or not). The only exception is [sta] which should be properly marked in the transcriptions.
  - Only noise deletions in the transcription were counted as wrong, not noise insertions.
  - The given transcription is given the benefit of the doubt; only obvious errors are corrected.
  - Errors were only determined on item level, not on word level
  - For speech a maximum of 5% of the validated items (=files) may contain a transcription error
  - For non-speech a maximum of 20% of the validated items (=files) may contain a transcription error.

References

Bernstein, J., Taussig, K., & Godfrey, J. (1994). Macrophone: An American English telephone speech corpus for the Polyphone project. In *Proceedings ICASSP-94*, Adelaide, pp. 81–83.

Cieri, C. (2006). What is quality? Proceedings Workshop "Quality assurance and quality measurement for language and speech resources". In *Proceedings LREC 2006*, Genova, Italy.

Cieri, C., & Liberman, M. (2000). Issues in corpus creation and distribution: The evolution of the linguistic data consortium. In *Proceedings LREC 2000*, Athens, pp. 49–56.

De Vriend, F., & Maltese, G. (2004). Exploring XML-based Technologies and procedures for quality evaluation from a real-life case perspective. In *Proceedings ICSLP-Interspeech 2004*, Jeju, Korea.

Den Os, E. A., Boogaart, T. I., Boves, L., & Klabbers, E. (1995). The Dutch Polyphone corpus. In *Proceedings Eurospeech 1995*, Madrid, Spain, pp. 825–828.

Dybkjaer, L., Hemsen, H., & Minkler, W. (Eds.) (2007). *Evaluation of text and speech systems*. Springer.

Fersøe, H. (2004). Validation Manual for lexicons. http://www.elra.info

Fersøe, H., & Monachini, M. (2004). ELRA validation methodology and standard promotion for linguistic resources. In *Proceedings LREC 2004*, Lisboa, pp. 941–944.

Gibbon, D., Moore, R., & Winski, R. (Eds.) (1997). *The EAGLES handbook of standards and resources for spoken language systems*. Mouton de Gruyter.

Höge, H., Draxler, C., van den Heuvel, H., Johansen, F. T., Sanders, E., & Tropf, H. S. (1999). Speechdat multilingual speech databases for teleservices: Across the finish line. In *Proceedings EUROSPEECH'99*, Budapest, Hungary, 5–9 Sep., pp. 2699–2702.

Höge, H., Tropf, H. S., Winski, R., Van den Heuvel, H., Haeb-Umbach, R., & Choukri, K. (1997). European speech databases for telephone applications. In *Proceedings ICASSP 97*, Munich, pp. 1771–1774.

Iskra, D., Grosskopf, B., Marasek, K., Van den Heuvel, H., Diehl, F., & Kiessling, A. (2002). SPEECON – Speech databases for consumer devices: database specification and validation. In *Proceedings LREC2002*, pp. 329–333.

Iskra, D., Siemund, R., Jamal Borno, J., Moreno, A., Emam, O., Choukri, K., Gedge, O., Tropf, H., Nogueiras, A., Zitouni, I., Tsopanoglou, A., & Fakotakis, N. (2004). OrienTel – Telephony databases across Northern Africa and the Middle East. In *Proceedings LREC 2004*, Lisbon, pp. 591–594.

Maegaard, B., Choukri, K., Calzolari, N., & Odijk, J. (2005). ELRA – European Language Resources Association – Background, recent developments and future perspectives. *Language Resources and Evaluation, 39*, 9–23.

Moreno, A., Choukri, K., Hall, P., Van den Heuvel, H., Sanders, E., & Tropf, H. (2004). Collection of SLR in the Asian-Pacific area. In *Proceedings LREC 2004*, Lisbon, Portugal, pp. 101–104.

Moreno, A., Comeyne, R., Haslam, K., Van den Heuvel., H., Horbach, S., & Micca, G. (2000b). SALA: SpeechDat across Latin America. Results of the first phase. In *Proceedings LREC 2000*, Athens, Greece, Vol. II, pp. 877–882.

Moreno, A., Lindberg, B., Draxler, C., Richard, G., Choukri, K., Euler, S., & Allen, J. (2000a). SpeechDat Car. A large speech database for automotive environments. In *Proceedings LREC 2000*, Athens, pp. 895–900.

Schiel, F., & Draxler, C. (2003). *The production and validation of speech corpora*. Bavarian Archive for Speech Signals. München: Bastard Verlag.

Shammass, S., & Van den Heuvel, H. (2002). Specification of validation criteria for lexicons for recognition and synthesis. LC-STAR, Technical report D6.1. (http://www.lc-star.com)

Simons, G., & Bird, S. (2003). The open language archives community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing, 18*, 117–128.

Strassel, S., Miller, D., Walker, K., & Cieri, C. (2003). Shared resources for robust speech-to-text technology. In *Proceedings EUROSPEECH 2003*, Geneva, pp. 1609–1612.

Van den Heuvel, H. (2004). Methodology for a Quick Quality Check (QQC). ELRA Technical report D1.2.

Van den Heuvel, H., Boves, L., & Sanders, E. (2000). Validation of content and quality of existing SLR: Overview and methodology. ELRA Technical report D1.1.

Van den Heuvel, H., Boudy, J., Bakcsi, Z., Cernocky, J., Galunov, V., Kochanina, J., Majewski, W., Pollak, P., Rusko, M., Sadowski, J., Staroniew, P., & Tropf, H. S. (2001). SpeechDat-E: Five Eastern European speech databases for voice-operated teleservices completed. In *Proceedings EUROSPEECH 2001*, Aalborg, Denmark, Vol. 3, pp. 2059–2062.

Van den Heuvel, H., Choukri, K., Gollan, C., Moreno, A., & Mostefa, D. (2006). TC-STAR: New language resources for ASR and SST purposes. In *Proceedings LREC 2006*, Genova, pp. 2570–2573.

Van den Heuvel, H., Choukri, K., Höge, H., Maegaard, B., Odijk, J., & Mapelli, V. (2003). Quality control of language resources at ELRA. In *Proceedings Eurospeech*, Geneva, Switzerland, pp. 1541–1544.

Van den Heuvel, H., Hall, P., Moreno, A., Rincon, A., & Senia, F. (2004a). SALA II across the finish line: A large collection of mobile telephone speech databases from North & Latin America completed. In *Proceedings LREC 2004*, Lisbon, Portugal, pp. 97–100.

Van den Heuvel, H., Höge, H., & Choukri, K. (2002). Give me a bug: A framework for a bug report service. In *Proceedings LREC2002*, Las Palmas, pp. 569–572.

Van den Heuvel, H., Iskra D., Sanders, E., De Vriend, F. (2004b). SLR validation: Current trends & developments. In *Proceedings LREC 2004*, Lisbon, Portugal, pp. 571–574.

Wittenburg, P., Broeder, D., Klein, W., Levinson, S., & Romary, L. (2006). Foundations of modern language resource archives. In *Proceedings LREC 2006*, Genova, pp. 625–628.